A fast algorithm for calculating an expected outbreak size on dynamic contagion networks

Epidemics 16 (2016) 56–62

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

A fast algorithm for calculating an expected outbreak size on dynamic contagion networks

Jessica Enright[a,*], Rowland R. Kao[b]

[a] Computing Science and Mathematics, University of Stirling, Stirling FK9 4LA, United Kingdom
[b] Boyd Orr Centre, Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow G61 1QH, United Kingdom

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form 12 May 2016
Accepted 18 May 2016
Available online 24 May 2016

Keywords:
Network modelling
Contagion on networks

ABSTRACT

Calculation of expected outbreak size of a simple contagion on a known contact network is a common and important epidemiological task, and is typically carried out by computationally intensive simulation. We describe an efficient exact method to calculate the expected outbreak size of a contagion on an outbreak-invariant network that is a directed and acyclic, allowing us to model all dynamically changing networks when contagion can only travel forward in time. We describe our algorithm and its use in pseudocode, as well as showing examples of its use on disease relevant, data-derived networks.

1. Introduction

Increasingly, models of contagion spread on highly structured populations are being used to inform disease control (Green et al., 2008; James et al., 2007; Eames et al., 2015; Danon et al., 2011). As these models become increasingly complex, simple and robust approaches to calculate the potential outbreak size become increasingly important. Here, we provide an algorithm which allows this to be done faster than current methods on a variety of real-world networks.

Previous research (Eames et al., 2015; Danon et al., 2011) uses simulation to estimate expected outbreak size on various different types of network. This is computationally intensive, and provides only stochastically derived estimates for the outbreak size. We present an exact method for calculating expected outbreak size. While our method does not apply to arbitrary contagion spread on general networks, which is a known NP-hard problem (Shapiro and Delgado-Eckert, 2012), it is relevant to the wide range of contagion examples that can be expressed on the well-studied class of directed acyclic graphs (DAGs): networks in which there are no directed cycles. This class is particularly useful for modelling temporally changing contact networks, and the notion that time (and therefore infection) only flows in one direction is central to our approach.

Our method has two advantages over simulation: it is computationally much faster, and it gives an exact answer rather than a statistical estimate. These two advantages are of particular importance in applications where a rapid estimate is important, without the requirement for a detailed behavioural or disease model, as in an outbreak situation with stringent externally-imposed timelines, or as an internal component in a larger software package that must complete a very large number of outbreak size calculations over a large number of different networks.

The method we describe here has much in common with several previously described methods: the novelty is largely in our algorithmic treatment and its use on a particular multi-layer directed acyclic graph (a structure also used in Kim and Anderson, 2012; Valdano et al., 2015) in order to incorporate a temporally changing network. We wish to highlight the relatedness of our approach to the methods of Rogers(Rogers, 2015), and Ludwig's method (Ludwig, 1975) as applied to a random network by House et al. (2012).

Rogers (2015) and Karrer and Newman (2010) describe the use of a cavity method on a network to calculate node risk and travel the development of an outbreak, as well as its final size. Rogers (2015) uses a tree approximation of a static network in its calculations of probability of given node's involvement in an outbreak; we apply a similar calculation to our directed acyclic graph.

Ludwig's method works on a system of pre-generated ranks in which nodes are assigned an order, and considered for infection in that order, and when applied to a network, requires the network be unchanged by an outbreak (Ludwig, 1975; House et al., 2012; Pellis et al., 2008). Given a starting node for the outbreak, nodes are

* Corresponding author.
E-mail address: jenright@gmail.com (J. Enright).

http://dx.doi.org/10.1016/j.epidem.2016.05.002

sorted by the length of their shortest paths to the starting node, with these shortest path lengths used as each node's "rank". Nodes are considered for infection by order of their ranks, with nodes closer to the starting node considered earlier. As described in House et al. (2012), an implementation of Sellke's construction (Sellke, 2012) on a network bears a close resemblance to Ludwig's method on a network.

As with Ludwig's method, we will consider nodes in a rough order of distance from an outbreak seeding set of nodes, though for our approach any topological ordering would suffice, and (again like in Ludwig's method) we will require our network to be invariant with respect to the outbreak.

We direct the reader to House et al. (2012) for a review of a wide variety of methods in use for calculating the probability mass function of a final outbreak size.

We present the method on a DAG derived from an infection network in Section 3. In Section 4 we show the method on several example networks, including two derived from real-world data. Section 5 compares our method to a series of simulations, demonstrating the advantages in speed and accuracy. We conclude with some adaptations which can be made to run the algorithm on more complex contagion networks, and some suggestions for further research. We provide open-source Python code which we hope will be of use in the future to other researchers.[1]

2. Overview of algorithm on a directed acyclic graph

We describe our approach in several steps: first, following Kim and Anderson (2012) we describe the production of a directed acyclic graph to describe a dynamically changing network. While we will focus on calculating on a dynamic network, it is possible to produce the directed acyclic graph required from a static network simply by repeating static contacts over many time steps.

Using this DAG as input, we then describe an efficient algorithm to calculate the expectation that any given node will be infected at a given time in an epidemic where individuals become immediately infectious and remain infections indefinitely (an SI model), or can recover and become immediately susceptible again (SIS). We allow an arbitrary choice, or distribution of choices, of starting nodes and times for the epidemic. Because expectations can be combined linearly (Hamming, 1991), this node-by-node expectation calculation enables us to calculate the expected size of an overall outbreak exactly at any fixed timepoint, again, either with a set starting node and time, or over a specified distribution of starting points.

2.1. Producing a directed acyclic graph from a dynamic network

In our preferred method for producing a DAG from a dynamic network, we essentially identify each agent at each time step with a node in the DAG, with an edge from one node (u, t) in the DAG to another $(v, t+1)$ if the state of the vertex u at time t can affect the state of the vertex v at time $t+1$. As in Kim and Anderson (2012) and Valdano et al. (2015) we use a multi-layered directed acyclic graph in which each layer is a time slice to encode a dynamically changing network of impulse edges. We assume throughout that disease cannot spread instantaneously, that is, an agent infected at t cannot infect another instantaneously, but is only able to infect others at $t+\epsilon$, where ϵ is an appropriately and arbitrarily small number. We also assume that the set of contacts that make up the network are known before the beginning of our calculation.

Let $G = (V, \vec{E})$ be a graph (or network) with vertices V and time-impulse directed edges $\vec{E}$. Let $\mathcal{T}$ be the relation between impulses and the times at which they occur. We assume that the range of $\mathcal{T}$ is a subset of the integers. Let $\mathcal{E}$ be the set of edges expressed as triples: (u, v, t) indicating an edge from u to v at time t, and let $Q : \mathcal{E} \to [0\ldots1]$ be the probability that, if the source of each impulse contact is infected, it will infect the destination of the edges.

Let V_T be the set: $\{(v, t)$ where $v \in V$, and $t \in [\min(range(\mathcal{T})) - 1 \ldots \max(range(\mathcal{T}))]\}$ Let $\vec{E_T}$ be the set:

- $\{((v, t) \to (u, s))$ where $u = v$ and $s = t+1\} \cup$
- $\{((v, t) \to (u, s))$ where $t = \mathcal{T}(u, v)$ and $s = t+1\}$

Let $\mathcal{P} : \vec{E_T} \to [0\ldots1]$ be a function from $\vec{E_T}$ to real-numbered probabilities between 0 and 1 such that:

- for edge $((u, t) \to (u, t+1))$, we set $\mathcal{P}(((v, t) \to (u, t+1)))$ to the probability that the disease persists at v from time t to $t+1$ and
- for edge $((v, t) \to (u, t+1))$ where $(v, u, t) \in \mathcal{T}$, we set $\mathcal{P}(((v, t) \to (u, s))) = Q(v, u, t)$

We have the building blocks of our directed acyclic graph in the form of a node set, an edge set, and probabilistic weights for the edges. Let graph $G_T = (V_T, \vec{E_T})$ be a directed graph: we know that G_T is acyclic because for every edge $((v, t) \to (u, s)) \in E_T$ we know that $s > t$; intuitively, the edges only go forward in time.

With the directed acyclic graph $G_T = (V_T, \vec{E_T})$ and the probability weighting function $\mathcal{P}$ we have the required input for our algorithm. Therefore, given a set of integer-time impulse contacts with probabilities of disease transmission associated with each contact, we can produce the graph we need, and use our algorithm to calculate expected outbreak size.

3. Expected outbreak size algorithm

While our algorithm below will work on any directed acyclic graph, we describe it in the context of a time-expanded graph as above, as this is the most relevant to our examples.

Let $G = (V, \vec{E})$ be a directed, acyclic graph as described above and $\mathcal{P} : \vec{E} \to [0\ldots1]$ be a function from the edges of G to probabilities such that $\mathcal{P}((u \to v))$ is the probability that u will infect v if it is, itself, infected. Note that, as described, there may be edges $((u, t) \to (v, t+1))$ where $u \neq v$ between different agents at successive times, as well as edges $((u, t) \to (u, t+1))$ between the same agent at successive times. The probability that an edge of the type $((u, t) \to (u, t+1))$ transmits is the probability that an infection of agent u at time t persists to time $t+1$. In general, the probabilities that edges transmit infection may differ: this is no impediment, so long as it is recorded in $\mathcal{P}$.

We start with a question: what is the expectation that (v, t) is infected in an epidemic with a known starting point (u, t_0)? If we consider all nodes at all times that could be infected in an epidemic stating at (u, t_0), we can identify the set of nodes that could directly infect (v, t): those that are the source of an edge leading into (v, t) that could, themselves, potentially be infected by an epidemic starting at (u, t_0). We call these the *parents* of (v, t), and due to the construction of the DAG we have used, we know that they are in time slice $t-1$. Let $A = \{(p_0, t-1), (p_1, t-1) \ldots (p_m, t-1)\}$ be the set of parents of (v, t) in a traversal of G from (u, t_0). Note that, if we are using a time-expanded graph as defined above, then exactly one p_i will be equal to u: exactly one parent of an agent at a time is that agent at the previous time. Then the probability that (v, t) is infected in an outbreak is the probability that at least one parent is infected and infects (v, t). Recall that $\mathcal{P}(((p_i, t-1) \to (v, t)))$

[1] https://github.com/magicicada/expected-outbreak-size.

is the probability that, if $(p_i, t-1)$ is infected, it infects (v, t). Then if $P_I((v, t))$ is the probability that (v, t) is infected,

$$P_I((v,t)) = 1 - [\prod_{i=0}^{m}(1 - P_I((p_i, t-1))]$$

To calculate this probability for (v, t) we need only know the probabilities that each of (v, t)'s parents are infected, and the probability of transmission along the inward edges from those parents: because the graph is directed and acyclic we can generate in linear time an ordering of the nodes in V such that every vertex occurs after all of its parents: for example, a topological ordering. Given this ordering we need only calculate the above probability for each node in turn. Then, for any given time, we can calculate the expected number of infected nodes that represent agents at that time: in this way we can produce an epidemic curve over time. We present this algorithm as pseudocode in Algorithm 1.

Note that a node that is not reachable from the initially infected set of nodes will have zero expectation of infection, and if every edge had a 100% probability of transmitting disease then every node reachable from the initially infected nodes would certainly eventually be infected. This is not usually the case, and so the set of nodes reachable from the initially infected nodes (such as might be computed using Dijkstra's algorithm), gives us an upper bound on the expected outbreak size calculated by the algorithm described here. Computing a topological (or pre-order) sort of the vertices in the directed acyclic graph takes time that is in the worst case linear in the size of that graph: in practice, we need only compute it for the vertices accessible from vertices in the outbreak seeding set, a significant time savings for many seeding sets, and do so using a simple breadth-first search. In fact, this computation can be done at the same time as the expectation values computation, and in practice therefore happens for free. If pre-computed, it takes time similar to the expectation computation traversal – on our testing machine about 0.08 ms for a network of 2699 nodes.

Algorithm 1. Algorithm to calculate the probability that each node in a time-expanded graph is infected in an epidemic at a given start node.

Input: A time-expanded graph $G = (V, \vec{E})$, a node n_{start} where the epidemic will be seeded, $\mathcal{P}$ a mapping of edges in G to probabilities that those edges will transmit infection if source of edge is infected
Output: A mapping from nodes in the time-expanded graph to probabilities that each is infected in an epidemic seeded at n_{start}
$T \leftarrow$ the DAG of all nodes of G reachable from n_{start}
$O \leftarrow$ a topological ordering of nodes in T
$\mathcal{D} \leftarrow$ an empty mapping
for $(u, t) \in O$ **do**
  $P \leftarrow$ parents of (u, t) in T
  probNotInfected $\leftarrow 1$
  for $(v, t') \in P$ **do**
    probInfectThis $\leftarrow \mathcal{P}[((v, t') \rightarrow (u, t)]$
    probNotInfected $\leftarrow$ probNotInfected $\cdot$ (1- probInfectThis $\cdot \mathcal{D}[(v, t')]$)
  end for
  $\mathcal{D}[(u, t)] \leftarrow$ 1- probNotInfected
end for
return $\mathcal{D}$

While we have described the intuition of our algorithm as for an epidemic with a single outbreak seed node in the network, it is equally applicable for an arbitrarily sized set of seeding nodes, and even an arbitrary probability of being seed node over any possible set of seed nodes – the calculation can be done in one traversal, and need not be repeated for each seed in the set. This will be especially valuable when an initial set of infected nodes is known, as is often the case in an outbreak that is discovered in progress.

4. Expected outbreak size calculation examples

To demonstrate the output and performance of the expectation calculation method we use a Python implementation to calculate the expected outbreak size of a SIS epidemic on two data-derived networks and a randomly-generated scale-free network and show the results in Fig. 1. In all our calculations we use a transmission probability of 0.4 for each contact, and a recovery probability of 0.2 at each time step. These values are not intended to represent any particular contagion, but rather are used as demonstration values. The units of our outbreak sizes are agent-timesteps, as we calculate the expectation that each agent is infected at each timestep, and then sum over all timesteps to find the overall expected agent-timestep epidemic size.

Our two data-derived networks are the network of cattle trades within Scotland in 2011 and message contacts on a Facebook-like social network for students at the University of California, Irvine (Opsahl and Panzarasa, 2009). Both datasets have temporally explicit directed impulse contacts. In these two data-derived networks, we use 10-day periods covering the entire network time period. We generated a random scale-free network using tools provided in the python `networkx` library, and assigned each edge three times chosen uniformly at random between time step 0 and 10, to be consistent with the ten-day time periods considered in the data-derived networks. The edge frequencies here are chosen solely for demonstration purposes, and result in a network with similar daily density to the cattle movement dataset. All calculations of expected epidemic sizes used over all the networks executed in a total of less than five minutes.

As is typical in real-world networks, many outbreaks are very small, and are not expected to leave their initial incursion point.

4.1. Cattle trading links in Scotland

While the majority of cattle movements between holdings in Great Britain are recorded, it has been possible in recent years for holdings to apply for linkages to other holdings that allow the owner of animals on that first holding to move those animals to and from the second holding without reporting. These linkages are a source of uncertainty in the cattle movement network in Scotland, as the magnitude and frequency of their use is unknown.

As a means of estimating these linkages potential impact on the size of an outbreak in Scotland, we used our rapid expected outbreak size calculation method to calculate the increase in expected outbreak size resulting from a simplified inclusion of these links, and to qualitatively examine the geographic distribution of holdings likely to be infected in a randomly seeded epidemic, comparing the situation with and without the inclusion of these links. The high level of uncertainty makes our technique especially appropriate: because we do not know the appropriate infection probabilities or relative risk of a contact, we must explore a very large parameter space, giving expected outbreak sizes at a large variety of disease and contact parameters. Such a large variety of parameters would require a computationally demanding number of simulation runs. The speed of our approach allowed us to produce a timely result. In addition, the expected outbreak size calculation gives farm-by-farm expectation of infection, allowing us to map the possible impact of these linkages at a farm level.

As we would expect, including linkages as movements increases expected outbreak size with higher probabilities of transmission and more frequent use of linkages increasing the expected outbreak size more (Fig. 2). Because of the uncertainly about the use of trade linkages, the results in Fig. 2 are shown over a variety of transmission settings and linkage uses, which, if computed by simulation, would have required a large number of simulations at a

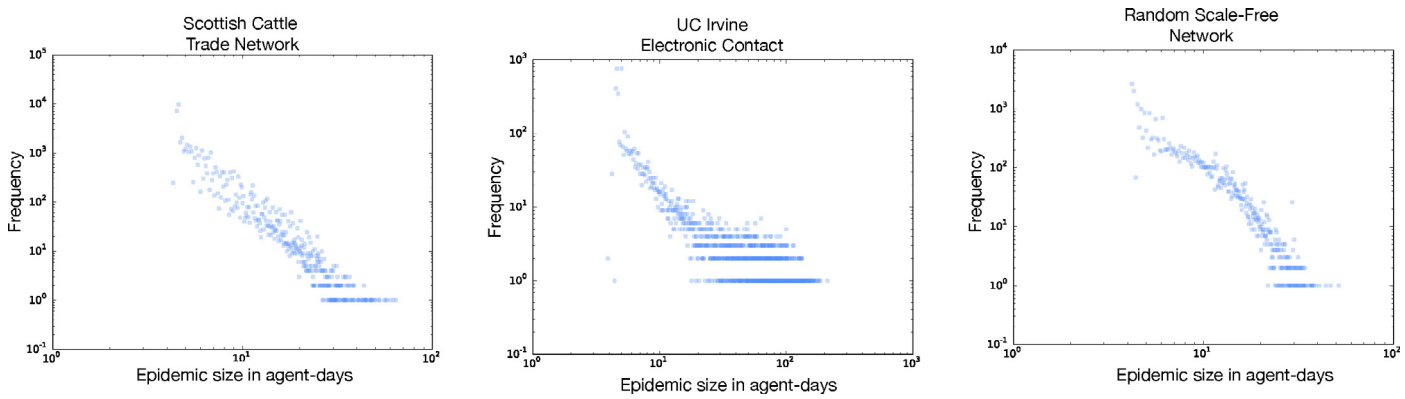

Fig. 1. Frequencies of expected outbreak sizes in agent time-steps started uniformly at random over all starting locations at the beginning of a 10 time-step period on three networks: on the left, a Scottish cattle trading network in 2011, in the middle a human electronic message contact network, and on the right a random scale-free network. We do not plot the sizes of outbreaks that are not expected to spread beyond their initial starting point.

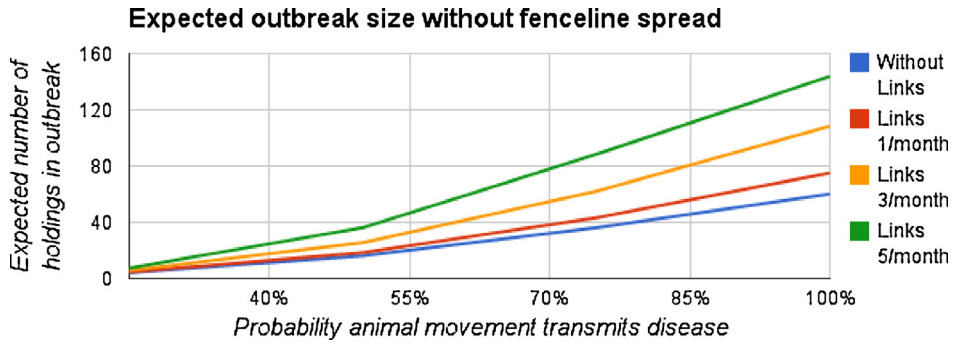

Fig. 2. The expected number of holdings infected in an outbreak of a disease spreading among cattle holdings in Scotland over a variety of probabilities of disease being transmitted by a batch movement of animals. We report expected outbreak sizes without registered linkages, and when each linkage is included once, three times, and five times per month. This calculation used animal movements from each month in 2013.

number of simulation settings, presenting a significant barrier to producing this result in a timely manner.

In addition to calculating the impact of linkages on the overall expected size of an outbreak, we also characterised the individual farms by their expectation of infection, which allows us to see, in Fig. 3, that the general geographic areas of most-likely infection remain the same with and without the linkages; a useful insight for geographic targeting of disease surveillance.

5. Simulation use compared to expected outbreak size calculation

Running both our algorithm and 100 strings of a standard forward micro-simulation approach, we calculated the expected outbreak size of a simulated SI contagion on the Scottish cattle movement network in January 2011 (with transmission probability of 0.4 with each contact, and the epidemic starting on the 1st of

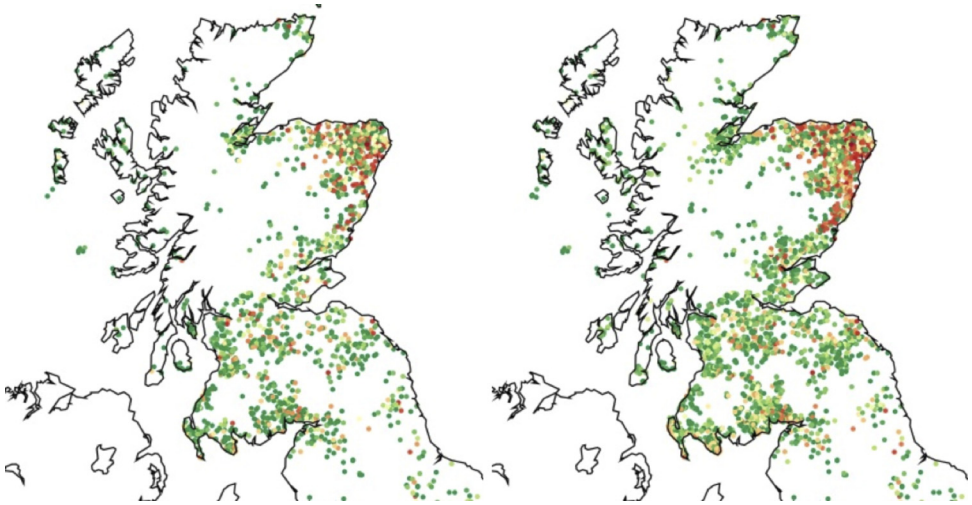

Fig. 3. Cattle holdings in Scotland, coloured by their probability of being infected in an outbreak seeded uniformly at random at a cattle holding in Scotland, with red being most likely and green being least likely. Holdings that are only expected to be infected by outbreaks seeded at that holding itself are omitted. On the left, outbreaks do not spread over linkages, on the right linkages are included in disease spread.

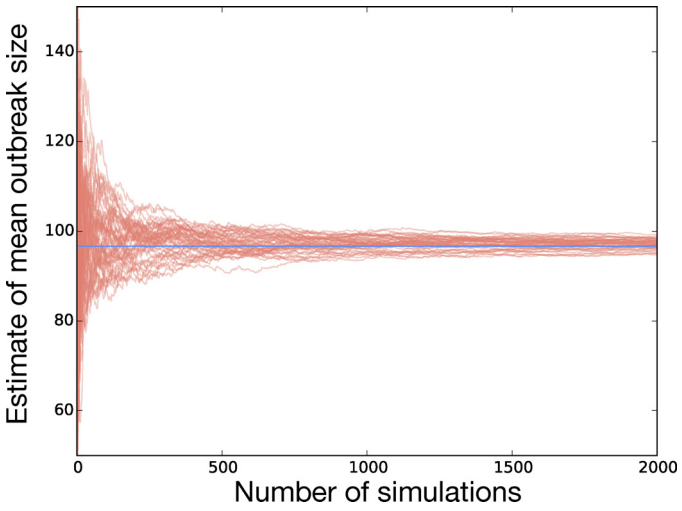

Fig. 4. The estimated outbreak size in agent-timesteps over the course of many strings of simulations on the Scottish cattle movement network in January 2011. Each serial string of simulations is one red line: values converge over a large number of simulations towards the expected outbreak size (shown in blue) as calculated using the method presented in this work. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

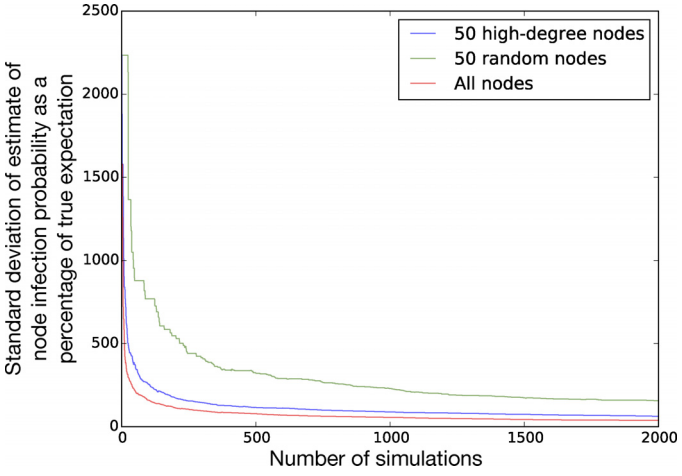

Fig. 5. The standard deviation of the simulation-generate estimate that a set of nodes will be infected as a percentage of the true expectation of that value over the course of many strings of simulations on the Scottish cattle movement network in January 2011 for all nodes, fifty randomly chosen nodes, and the fifty nodes with the highest number of contacts over the entire month.

January, located uniformly at random across all cattle holdings in the network). Each simulation string consists of a series of independent micro-simulations, with the estimate of expected outbreak size after some set of simulations calculated as the mean of the outbreak sizes in the simulations to that point. The results are presented in Fig. 4, where the blue line represents the output of our algorithm, which is the *exact* expectation of the size of the outbreak on the network. We see that the simulation method can take many iterations to converge to anything close to the exact answer – and even then there is some error.

In the course of our calculation, we find the expectation that any given node will be infected in an outbreak started across a uniform distribution of starting points for the epidemic. As for the overall expected outbreak size, the simulated value converges over a large number of simulations towards the expected value. The rate of this convergence is not uniform over all sets of nodes; in Fig. 5 we show this convergence as the standard deviation of the simulated values as a percentage of the true expected value for all nodes, fifty

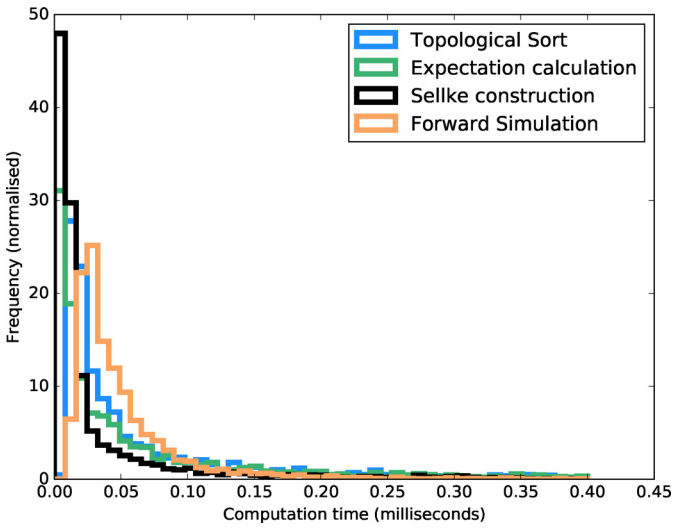

Fig. 6. Frequencies of computation times for a single run of the topological sort required for our expectation calculation, our exact expectation calculation, simulation using the Sellke construction, and simple forward simulation. All simulations were run on a late 2013 iMac with a 3.2 GHz Intel Core i5 processor, and were run on the Scottish cattle movement network in January 2011, with outbreaks seeded at each possible holding on January 1st. The mean computation time for the topological sort was 0.08 ms, for the expectation calculation was 0.08 ms, for the Sellke construction was 0.05 ms, and for the forward simulation was 0.06 ms.

randomly chosen nodes, and the fifty nodes with the highest number of contacts over the entire month. As expected, the convergence is slowest and poorest for fifty random nodes, better and faster for the fifty nodes with the most contacts, and best for all nodes overall. The inequality in speed of convergence reminds us that caution should be used when interpreting simulation-derived risk values for individual agents, as some estimates may be far more reliable than others, with poorly connected agents being particularly susceptible.

We have compared the wall-clock running times for the topological sort required by our algorithm, our main expectation calculation, the basic forward simulation, and a simulation using the Sellke method for networks, as outlined and implemented in House et al. (2012) in Fig. 6. We find similar distributions of running times over 3000 instances of each computation, with similar mean computation times: our expectation calculation takes on average under twice the time required for a Sellke construction simulation or forward micro-simulation, but as in Fig. 4, many simulations would be required to converge to the expected value. Note that the network used is fairly large (approximately 2700 nodes) and very sparse, so we have made adaptations to the Sellke method on networks described in House et al. (2012) to avoid both matrix operations that include the entire adjacency matrix of the network and calculation of unneeded pseudo-random numbers. On a smaller, denser, network, we would expect the Sellke construction to be significantly faster than the forward simulation.

Our directed acyclic graph construction results in a graph that is larger when more time steps must be considered. This means that a longer time period or smaller time steps will result in a larger graph, and therefore slower running time: this relationship, however, is approximately linear: the processing becomes slower in proportion to the number of additional nodes and edges. If the time-span is very long, with small time steps and dense edges, the expectation calculation may become unfeasible.

A number of algorithmic modifications can be made to mitigate this effect when the contagion is simple and the contact network is sparse: for example, if a node v has no contact with other notes between times t_i and t_j, then we can replace the node copies of (v, t_k) for $i \leq k \leq j$ with a single edge from (v, t_i) to (v, t_k) with an

appropriately modified probability of persistence as the weight on that edge. In a sparse dynamic network, this will significantly cut down the number of nodes in our directed acyclic expansion.

In addition to calculating expected outbreak size, or time to total infection, we can measure the contribution of a single node, edge, or set of nodes or edges, to the expected outbreak size. The simplest way to do this is to calculate the expected outbreak size with the node, nodes, or edges in the graph, and then with them removed from the graph. The difference between these sizes could then serve as a metric of the contribution of that node or edge set. For any constant k, we can then calculate the optimal node or edge set of size k to remove to minimise expected outbreak size in an n^k (if a vertex set) or m^k (if an edge set) multiple of the running time of expected outbreak size algorithm. We call the amount by which the removal of a set of vertices or edges decreases the expected outbreak size that set's *removal difference*.

6. Adaptations

While we have described the production of a time-expanded graph from impulse contacts and a calculation for an SI or SIS contagion, adaptations can be made to the vertex and edge sets of the time-slice network we have described so as to approximate different types of contacts and contagions, while using the same general idea of calculating an expected outbreak size. We outline several of them here, but this list is by no means exhaustive.

6.1. Non-impulse edges

The simplest approximate adaptation for edges that exist continuously over a time period is to add a number of impulse edges separated by the minimum time step in the dynamic network, with transmission probabilities adjusted appropriately. Then one can produce the directed acyclic graph as described above. While this approach requires discretisation, so does any other approach that requires computer processing. This could also be used when different types of contacts persist for different periods of time: for example, consider the differing length of contacts between cattle herds at two farms that physically share a fenceline, as compared to the directed and instantaneous contact due to a trade, or the sharing of equipment.

While we have described the production of a directed, acyclic graph from temporally explicit network contacts, one could also include persistent contacts: for example, fenceline adjacency contacts between farms. The main disadvantage of doing so is an increase in the density of the DAG, and therefore a corresponding increase in the size of breadth-first search trees from an outbreak seed set, and an increase in running time. This occurs in its most extreme form if we add a fenceline contact edge between every farm and its neighbours at every time step. While this does not change the asymptotic time complexity of the algorithm as a function of the number of edges, it slows computation considerably if the network would otherwise be sparse. For example: when running our example computation on cattle trades, if we include the true fenceline adjacency edges in our computation (which increases the number of edges in the DAG by a factor of 229), the mean computation time for an outbreak seeded at a single source increases from 0.08 to 270 ms: if a calculation over a long time frame is required, the exact calculation may no longer be appropriate. Mean forward-simulation computation time is also increased by the addition of these edges, increasing from 0.05 ms to 3900 ms, with most of this additional time being spent in generating random numbers. However, note that the computation time of a forward-simulation is a function of the probability that edges transmit infection: a large epidemic takes far longer to simulate than one that dies out quickly. This is less true for the exact calculation method: if there is even a very small expectation of infection at a node, we will continue to compute for that node's children.

6.2. Adapted for undirected edges

Undirected edges between vertices can be modelled as two directed edges; because these edges will go forward in time, they will not cause a cycle in the time-expanded graph.

6.3. Latent infection period

Many infections e.g. Reynolds (2006) have a latent infection period: a period of time after an agent is infected when it is not yet infectious. This can be approximated by changing the forward contact arrows in the time-expanded graph so that, if δ_t is the latent infectious period, then contact edges between different agents are added between (u, t) and $(v, t + \delta_t)$, rather than between (u, t) and $(v, t + 1)$.

6.4. Distributions of outbreak sizes

In the form described above, our expectation calculation gives only a single expected outbreak size for a given initial seed of expectations. However, the combination of the DAG produced, the seed of expectations, and the probabilities of transmission and recoveries stored on the edges of the graph contains all the information required to reconstruct the distribution of outbreak sizes that would be given by a stochastic forward simulation (though to do so fully would require computation equivalent to full forward simulation). Intermediate levels of detail about the distribution of expected outbreak sizes is possible by tuning the set of initial seeds for the epidemic. For example: perhaps an initially infected node has a neighbour that is of high importance, and whether the outbreak is large or small depends largely on whether that important neighbour is infected. We could run our expectation calculation with that important neighbour specified as an initially infected seed, and then again with our that neighbour. Thus with additional specified seedings, we can reconstruct more and more of the distribution of outbreak sizes; however, if this is to be done efficiently, we would need a pre-existing correct idea of which vertices are likely to be important. If full detail of outbreak size distribution and variance is required, a different approach may be more appropriate.

7. Concluding remarks

We have presented our use of a traversal-based algorithm to calculate expected outbreak size on a temporally-changing network, shown several examples of its use for rapid outbreak size estimation, and described its use for identifying epidemiologically important nodes in a network. We have demonstrated that the algorithm is practical, and has distinct advantages over simulation.

While we have investigated only the simplest contagion examples here to demonstrate the applicability of the approach, in principle it can be extended to considerably more intricate scenarios, so long as they do not violate the fundamental assumptions of the approach (for example, the algorithm must operate on a contact DAG, invariant to any disease outbreak on it, with known edge transmission probabilities). We suggest several areas for future investigation:

1. We have suggested several adaptations of our method to include non-impulse edges, a latent infection period, etc. Can this approach be adapted to include removal or immunity of a node after recovery?

2. We (and Kim and Anderson, 2012) suggested the use of our approach for calculating epidemiologically-important sets of vertices: on which networks will this approach improve on simply removing the vertices with highest degree? Can we characterise networks on which this will be the case in terms of simple network parameters?

We have made a python implementation of our algorithm available on GitHub, and hope that it will become a useful tool to researchers who might otherwise calculate outbreak size by simulation.

The implementation of such fast algorithms can be particularly important where time-bound outcomes are important – for example, the application involving cattle trading links in Scotland was developed in the context of policy-driven questions with short, externally driven deadlines, where a fully customizable simulation may not have been feasible or appropriate.

Acknowledgement

The authors gratefully acknowledge funding from the Scottish Government as part of EPIC: Scotland's Centre of Expertise on Animal Disease Outbreaks.

References

Danon, L., Ford, A.P., House, T.A., Jewell, C.P., Keeling, M.J., Roberts, G.O., Ross, J.V., Vernon, M.C., 2011. Networks and the epidemiology of infectious disease. In: Interdisciplinary Perspectives on Infectious Diseases (Article No. 284909).

Eames, K., Bansal, S., Frost, S., Riley, S., 2015. Six challenges in measuring contact networks for use in modelling. Epidemics 10 (0), 72–77 (Challenges in Modelling Infectious Disease Dynamics).

Green, D.M., Kiss, I.Z., Mitchell, A.P., Kao, R.R., 2008. Estimates for local and movement-based transmission of bovine tuberculosis in British cattle. Proc. R. Soc. Lond. B: Biol. Sci. 275 (1638), 1001–1005.

Hamming, R.W., 1991. The Art of Probability: For Scientists and Engineers. Advanced Book Classics. Perseus Books.

House, T., Ross, J.V., Sirl, D., 2012. How big is an outbreak likely to be? methods for epidemic final-size calculation. Proc. R. Soc. Lond. A: Math. Phys. Eng. Sci. 469 (2150).

James, A., Pitchford, J.W., Plank, M.J., 2007. An event-based model of superspreading in epidemics. Proc. R. Soc. Lond. B: Biol. Sci. 274 (1610), 741–747.

Karrer, B., Newman, M.E.J., 2010. Message passing approach for general epidemic models. Phys. Rev. E 82 (July), 016101.

Kim, H., Anderson, R., 2012. Temporal node centrality in complex networks. Phys. Rev. E 85 (February), 026107.

Ludwig, D., 1975. Final size distribution for epidemics. Math. Biosci. 23 (1), 33–46.

Opsahl, T., Panzarasa, P., 2009. Clustering in weighted networks. Soc. Netw. 31 (2), 155–163.

Pellis, L., Ferguson, N.M., Fraser, C., 2008. The relationship between real-time and discrete-generation models of epidemic spread. Math. Biosci. 216 (1), 63–70.

Reynolds, D., 2006. A review of tuberculosis science and policy in Great Britain. Vet. Microbiol. 112 (24), 119–126 (4th International Conference on *Mycobacterium bovis*).

Rogers, T., 2015. Assessing node risk and vulnerability in epidemics on networks. Europhys. Lett. 109 (2), 28005.

Sellke, T., 1983. On the asymptotic distribution of the size of a stochastic epidemic. J. Appl. Probabil. 20 (2), 390–394.

Shapiro, M., Delgado-Eckert, E., 2012. Finding the probability of infection in an sir network is np-hard. Math. Biosci. 240 (2), 77–84.

Valdano, E., Ferreri, L., Poletto, C., Colizza, V., 2015. Analytical computation of the epidemic threshold on temporal networks. Phys. Rev. X 5 (April), 021005.